Focused Retrieval Using Topical Language and Structure

A.M. Kaptein

Archives and Information Studies, University of Amsterdam
Turfdraagsterpad 9, 1012 XT Amsterdam, The Netherlands
a.m.kaptein@uva.nl

Abstract

We investigate focused retrieval techniques that deal with the increasing amount of structure on the web. Our approach is to combine multiple representations of web information in a common framework based on statistical language models. In this framework, it will be possible to derive a topical language model of the actual language-use on web pages on a certain topic—such as arts, business, entertainment, education, etc.—using the unigrams and bigrams taken from the plain text of the web pages. Similarly, it will be possible to derive models of the structure of web pages to distinguish between blogs, FAQs, personal web pages, etc. Structural characteristics of a web page include, amongst others, tagname statistics and parent-child tags. We will build a multiple level language model to exploit the information contained in the topical language and structure models. The .GOV2 corpus will be used as a test collection on which queries will be run on different topical categories and on web pages with different structures. We plan to develop so-called parsimonious models to derive a compact representation and to handle dependencies between representations of the data.

Keywords: Focused Retrieval, Web Retrieval, Language Modeling, Relevance Feedback

1. Background and related work

The Internet gives us access to an unprecedented amount of data and information. This creates great challenges and opportunities. As to challenges, the growth of information prompts the need for methods that can help to organize, cluster and classify information, and improve ways of accessing it. This research addresses robust and focused retrieval techniques that deal with the increasing amount of structure on the web. Whereas part of this is orchestrated, for example by the adoption of Semantic Web techniques [Berners-Lee et al., 2001], much of it is due to the self-organizing nature of the Web [Kumar et al., 1999]. As a result, the structure on the web is rich but highly heterogeneous and noisy, and there is a need for techniques that help organize noisy structured information. Important initial steps have been taken here. Craswell et al. [2001] highlighted the importance of compact document representations based on anchor-texts for the specialized task of home page finding. For the same home page finding task, Kraaij et al. [2002] showed the importance of URL information, and introduced mixture language models that incorporate anchor text and URL information. This approach was extended by Ogilvie and Callan [2003a], addressing now more general known-item searches. Finally, using similar approaches, Kamps et al. [2004] and Kamps [2005] addressed various URL and link priors for both known-item search and topic distillation. Language models have been used for structured data in several other ways. Ogilvie and Callan [2003b] applies techniques from stochastic context-free parsing to XML retrieval. In related work outside the language modeling framework, Craswell et al. [2005] address a range of query-independence evidence for BM25. They propose transformations that take into account dependencies between the feature values and the content-scores.

The term language models originates from probabilistic models of language generation developed for automatic speech recognition systems in the early 1980s (see e.g. [Rabiner, 1990]), but at that time they had already been in existence for about 70 years. Language models have had quite an impact on information retrieval research. Essential in language modeling approaches is so-called smoothing. Smoothing is the task of re-evaluating the probabilities, in order to assign some non-zero probability to query terms that do not occur in a document. As such, smoothed probability estimation is an alternative for maximum likelihood estimation. The standard language modeling approach to information retrieval uses a linear interpolation smoothing of the document model with a general collection model (see e.g. [Berger, 1999; Song, 1999]). There are other approaches to smoothing language models (see e.g. [Jelinek, 1997]), some of which have been suggested for information

retrieval as well, for instance smoothing using the geometric mean and backing-off by Ponte and Croft [1998]; and Dirichlet smoothing and absolute discounting suggested in a study by Zhai and Lafferty [2001]. In a sense, all smoothing approaches somehow combine two representations of the data (a document model and a collection model) into a new representation. We plan to use similar approaches to combine many more document representations into a single representation.

2. Description of proposed research

Large-scale, general purpose web search engines (most notably Google) have been quite successful in keeping up with the size of the web by adding new sources of information to traditional full-text indexes: for instance anchor texts and hyperlink structure. However, the web is growing out of reach of these techniques, and companies like Google have started to offer specialized web search services focusing on for instance internet shopping (Froogle) and scientific documents (Google Scholar inspired by CiteSeer [2005]). Such services use lay-out analysis techniques and simple information extraction techniques that exploit web conventions—extracting e.g., product names, prices, author names, etc.—to come up with domain-dependent structured document representations that are far more complex than the full-text/anchor text/link structure representations. As such, those services provide very accurate and focused search.

However, specialized search engines only provide focused search if the user is first able to find the specialized search engine of his/her choice, if one that caters for the user's problem exists at all. Whereas specialized search engines are part of the answer to the increasing size of the web—they identify more structured information and provide more focused search—they also introduce new information overload problems. We believe we can have the best from both worlds: very focused and accurate search without the need for the user to preselect the domain beforehand.

Our main research problem is the following:

- Can we provide accurate and focused search by adding structure and combining multiple, complex representations within a common information retrieval modeling framework using generic web retrieval models?

There is an increasing amount of structure in documents on the web. Like the specialized search engines mentioned above, we intend to develop structured document representations, where the structure is derived from the document text, the document structure, the URL, the hyperlink structure, time, geographic location, text classification, manually assigned metadata, and more. All these sources of evidence can provide crucial retrieval cues. We will start by focusing on structure derived from the document text and the document structure. Hence, our first research question is:

- Can we use topical language and topical structure similarity to improve retrieval results?

Topical language similarity

We will investigate the particular language use for specific topics and types of web pages, and build localized models for them. For example, we can consider the type of language used on educational web pages or on home pages. We can anchor the topical content types on Internet directories as provided by the Open Directory, Yahoo!, Google, or on-line encyclopedias such as Wikipedia [Rode, 2004]. Particular language usage provides a layer of topical language models that can be incorporated in the language modeling framework.

Topical structure similarity

Whereas a typical search engine will disregard almost all of the document markup, and focus primarily on the textual content presented to the user, we expect that there are valuable retrieval cues in structural aspects of web pages. This is highly related to the emergence of web conventions, in which specific types of web content have adopted a similar look-and-feel. As such there is great structural similarity between, for example, home pages of people, on-line product pages, FAQs, blogs, etc. Abstracting the structure of pages provides another layer of structural language models that can be incorporated in the modeling framework.

Besides using the information that we can find in the document collection, we also want to use the information that can be provided by the user implicitly or explicitly. A bottle-neck for providing more focused retrieval is in the shallowness on the client-side, i.e. users who provide no more than a few keywords to express their complex information needs. We want to use implicit and explicit feedback info to improve and to group the retrieval results. Therefore, our second research question is:

- Can we use topical language and structure similarity in combination with implicit and explicit feedback to enhance the user's search experience?

Implicit information about the users geographical location might for instance come from his/her IP number, or implicit user preferences might be derived from click-through information. Explicit feedback can be obtained in two ways. It can be in the form of a suggestion that is relevant to the given query, e.g. "Do you want to focus on documents from around 11 September?" or "Are you looking for a person's home page?". Using this feedback a new ranked list of results will be provided. The second option is to cluster the retrieved results into the predefined topical and structural categories.

3. Outline of Models

Our modeling framework will be based in so-called statistical language models for information retrieval [Hiemstra, 2001]. The basic retrieval model has been successfully combined with models of non-content information that use for instance link information and URL-type in web retrieval [Hiemstra and Kraaij, 2005; Kamps et al, 2004; Kamps, 2005; Kraaij et al., 2002]. Our approach is to go beyond the "document as a bag-of-words" models by bringing more and more sources of evidence into the models and to extend existing language modeling approaches in several ways.

- *Relevance models*

Our approach to implicit feedback is related to relevance models [Lavrenko and Croft, 2003] in which the set of initially retrieved documents is included in the model as a layer between the document and the collection model. For example, we believe that there is great potential in the combination of metadata (either in the form of e.g., the Yahoo! directory, or metadata deduced from previous queries) with derived representations. Using such metadata, we will be able to derive topical and structural models, i.e., models of the language typically used in documents on a certain topic. A model built from documents that were assessed as relevant to a query that is similar to the new query can be used to provide more targeted search. Topical language and structure models can also be used to classify (initially retrieved) documents using text categorization techniques. This, in turn, provides crucial information for adjusting smoothing parameters, for result clustering, or for asking follow-up questions like "Do you want to focus on homepages?"

- *Parsimonious language models*

Usually, each language model representation is defined independently from the others and they are combined later on. However, independent definitions of language modeling components lead to large, redundant combined representations. To model structurally complex document representations, we need a method that rewards parsimony. Parsimonious models [Sparck-Jones et al., 2003] explicitly address the relation between several representations of the document. As such, they effectively model for each partial representation what it adds to the document representation as a whole, thus avoiding redundancy.

It has recently been shown that parsimonious models improve retrieval performance in both text search [Hiemstra et al., 2004] and image search [Westerveld and De Vries, 2004]. Techniques from parsimonious language models allow us to define those representations in a layered fashion by explicitly addressing dependencies between document representations. As additional but important bonus, avoiding redundancy results in significantly smaller models. This leads to a reduction of storage space, and a reduction of query processing time—two key requirements for effective large-scale web retrieval.

4. Research methodology and proposed experiments

We will start by investigating whether we can exploit topical language and structural similarity of web pages in a certain topic category or with a certain structure to improve retrieval results. We expect topical language models will work best for topical categories, i.e. if the topic category is 'Education' words like 'school' and 'course' are likely to occur. Also we expect structural similarity to be a good indicator of the structural type of a web page, i.e. a home page contains maybe a photo or a logo and some links to other pages. Or maybe it is the combination of topical language and structural similarity that produces the best results.

We will investigate also if it is possible to recognize a structural type of webpage by using a language model that takes in plain text. For instance, words like 'welcome' and 'home page' could indicate the structural type of this page is 'Homepage'. Similarly it could be possible to recognize a topical category by certain structural characteristics.

Up to this moment, we have analyzed the type of documents in the .GOV2 corpus and queries used in previous TREC Terabyte tracks to specify a set of topic categories and a set of structural types of web pages. For practical reasons, the .GOV2 corpus will be used as the initial test collection for our experiments. To be able to create our models we need queries that fall within our topic categories and that find pages of different structural types of web pages. We have defined 13 topic categories, including some combinations of categories.

Examples of topic categories are:

- Education
- Health care
- Transportation
- Environmental issues
- Human rights
- Etc.

We will reuse as many queries as possible from previous Terabyte tracks and of the Million Query track. For each topic category we can find 1 to 12 queries that have been assessed in previous Terabyte tracks and a query produces on average 200 relevant results. In total, our test set contains 8986 relevant web pages. However, the grouping of the relevant results does not cover the full range of topics that fall within a topic category for most of our topic categories.

Besides topical categories, we also started looking at different structural types of webpages.

Some examples of structural types of web pages (not restricted to the .GOV2 corpus) are:

- Personal homepage
- Organizational homepage
- How to / FAQ page
- Discussion / Forum
- Hub (page with links)
- Scientific papers
- Online shop
- Blogs
- Etc.

Unfortunately, the .GOV2 corpus does not represent most of the above structural types. From past TREC results the only type that we can retrieve and analyze is homepages. Besides reusing queries we will therefore also create and assess a number of queries ourselves that cover each topic category and structural type of web page that is represented in the .GOV2 corpus. Another approach to find documents representative of a topic category to build a topical language model could be to take documents from an Internet directory such as Yahoo! or Wikipedia.

Using the results from the reused and our own queries we can create topical language models using unigrams and bigrams taken from the plain text of the web pages. The structure of a web page is a bit more difficult to capture in a model. We can use the following characteristics to represent the structure of a web page:

- use tagname statistics (like unigrams)
- use parent-child tags (like bigrams)
- use unlabeled graph and tree-similarity
- categorize html-tags into more abstract semantic labels (image, section, paragraph, table)

Once we have created a topical or a structural model we can use it to score either the query or the retrieved documents. The model with the highest score can be used as the extra layer in the language model to get a better ranking. We can use leave-one-out evaluation to determine whether our model is an improvement to the standard language model.

5. Discussion

Our research is still in an early stage, and we have only started to explore possible approaches to building more informed information retrieval models. Hence, we are open to any advice or comments on the current approach, or on alternatives and extensions of the approach. Specifically, there are a number of open issues that we would like to get feedback on:

- The standard TREC test collection does not suit all our purposes. Are there other sources of training and test data that could be relevant for our project?
- We plan to develop topics covering various topic categories and structural types. What would be a good way to create topics and how can we structure the assessment process?
- We are very interested in the user's point of view. Are there related user studies in the literature? Can we use some of the research on implicit feedback?

6. Acknowledgements

This research is done in the framework of the EfFoRT (Effective Focused Retrieval Techniques) project. Other members of the research team include Dr.ir. Jaap Kamps, Dr.ir. Djoerd Hiemstra and MSc. Rongmei Li. This research is supported by the Netherlands Organization for Scientific Research (NWO, grant # 612-066-513).

REFERENCES

- Berger, A. and Lafferty, J. (1999) Information retrieval as statistical translation. *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 222–229.
- Berners-Lee, T. , Hendler, T. and Lassila, T. (2001) The semantic web. *Scientific American*, **284**(3), 34–43.
- CiteSeer. (2005) CiteSeer: Scientific literature digital library. <http://citeseer.ist.psu.edu/>.
- Craswell, N., Hawking, D. and Robertson, S. (2001) Effective site finding using link anchor information. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, USA.
- Craswell, N., Robertson, S., Zaragoza, H. and Taylor, M. (2005). Relevance weighting for query independent evidence. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*, pp. 416–423. ACM Press, New York, USA.
- Hiemstra, D. (2001) *Using language models for information retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente.
- Hiemstra, D. (2001) *Using language models for information retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.
- Hiemstra, D., Robertson, S. and Zaragoza, H. (2004) Parsimonious language models for information retrieval. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178–185. ACM Press, New York, USA.
- Hiemstra, D. and Kraaij, W. (2005) A language modeling approach to TREC. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press.
- Jelinek, F. (1997) *Statistical Methods for Speech Recognition*. MIT Press.
- Kamps, J., Mishne, G. and De Rijke, M. (2004) Language models for searching in Web corpora. *NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*.
- Kamps, J. (2005) Web-centric language models. *Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management (CIKM 2005)*. ACM Press, New York, USA.
- Kraaij, W., Westerveld, T., and Hiemstra, D. (2002) The importance of prior probabilities for entry page search. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 27–34. ACM Press, New York, USA.

- Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999) Trawling the web for emerging cyber-communities. *Proceedings of the 8th International World-wide web conference WWW8*, pp. 403–415. Elsevier Science, Amsterdam.
- Lavrenko, V. and Croft, W. (2003) Relevance models in information retrieval. In Croft, W. and Lafferty, J. (eds), *Language Modeling for Information Retrieval*, pp. 11–56. Kluwer Academic Publishers.
- Ogilvie, P. and Callan, J. (2003a) Combining document representations for known-item search. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 143–150. ACM Press, New York, USA.
- Ogilvie, P. and Callan, J. (2003b) Language models and structured document retrieval. *Proceedings of the first workshop on the evaluation of XML retrieval (INEX)*, pp. 33–40.
- Ponte, J. and Croft, W. (1998) A language modeling approach to information retrieval. *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*, pp. 275–281.
- Rabiner, L. (1990) A tutorial on hidden markov models and selected applications in speech recognition. In Waibel, A. and Lee, K. (eds.), *Readings in speech recognition*, pp. 267–296. Morgan Kaufmann.
- Rode, H. and Hiemstra, D. (2004) Conceptual language models for context-aware text retrieval. *Proceedings of the 13th Text Retrieval Conference (TREC)*.
- Song, F. and Croft, W. (1999) A general language model for information retrieval. *Proceedings of the 8th International Conference on Information and Knowledge Management, CIKM'99*, pp. 316–321.
- Sparck-Jones, K., Robertson, S., Hiemstra, D. and Zaragoza, H. (2003) Language modelling and relevance. In Croft, W. and Lafferty, J. (eds), *Language Modeling for Information Retrieval*, pp. 57–71. Kluwer Academic Publishers.
- Westerveld, T. and De Vries, A. (2004) Multimedia retrieval using multiple examples. *International Conference on Image and Video Retrieval (CIVR'04)*.
- Zhai, C. and Lafferty, J. (2001) Model-based feedback in the language modeling approach to information retrieval. *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)*, pp. 403–410.